

EXHIBIT V

**STATISTICAL METHODS
IN
CANCER RESEARCH**

**Volume 1 – THE ANALYSIS OF
CASE-CONTROL STUDIES**

INTERNATIONAL AGENCY FOR RESEARCH ON CANCER

WORLD HEALTH ORGANIZATION



INTERNATIONAL AGENCY FOR RESEARCH ON CANCER

STATISTICAL METHODS IN CANCER RESEARCH

VOLUME 1 - The analysis of case-control studies

BY

N.E. BRESLOW & N.E. DAY

TECHNICAL EDITOR FOR IARC

W. DAVIS

IARC Scientific Publications No. 32

INTERNATIONAL AGENCY FOR RESEARCH ON CANCER

LYON

1980

The International Agency for Research on Cancer (IARC) was established in 1965 by the World Health Assembly as an independently financed organization within the framework of the World Health Organization. The headquarters of the Agency are at Lyon, France.

The Agency conducts a programme of research concentrating particularly on the epidemiology of cancer and the study of potential carcinogens in the human environment. Its field studies are supplemented by biological and chemical research carried out in the Agency's laboratories in Lyon and, through collaborative research agreements, in national research institutions in many countries. The Agency also conducts a programme for the education and training of personnel for cancer research.

The publications of the Agency are intended to contribute to the dissemination of authoritative information on different aspects of cancer research.

First reimpression, 1982
Second reimpression, 1983
Third reimpression, 1989
Fourth reimpression, 1990
Fifth reimpression, 1992
Sixth reimpression, 1994
Seventh reimpression, 1998
Eighth reimpression, 2000

ISBN 92 832 0132 9

International Agency for Research on Cancer 1980

The authors alone are responsible for the views expressed in the signed articles in this publication.

REPRINTED IN THE UNITED KINGDOM

However, using expression (3.1) and the balance in the design one can show that the odds ratio in the pooled table, ψ_p , lies between unity and ψ , that is we have:

$$1 < \psi_p < \psi, \text{ if } \psi > 1$$

or

$$1 > \psi_p > \psi, \text{ if } \psi < 1.$$

Thus, in contrast to the unbalanced situation, where the confounding effects can be either positive or negative and where the pooled odds ratio can be the opposite side of unity from the within-stratum odds ratio, *with a balanced design the expected pooled odds ratio is both on the same side of and closer to unity than the expected within-stratum odds ratio*. An unstratified analysis will bias the odds ratio towards unity, unless the confounding and exposure variables are (conditional on disease status) independent, but not change the side of unity on which the odds ratio lies.

Obviously the odds ratio of C with disease bears no relation to the true association. *If a factor has been balanced, the data so generated give no information on the association of that factor with disease*. Interaction can still be estimated, however, as is discussed in § 3.5.

Balance, as described above, where the control series is chosen to ensure equal frequency of cases and controls in different strata, is sometimes referred to as frequency matching. On other occasions, where for each case a set of controls is chosen to have the same, or nearly the same, values of prescribed covariates, we speak of individual matching. In later chapters matching refers specifically to individual matching.

Incorporation of matching factors in the analysis

The purpose of matching, as we have just seen, is to control confounding and increase the information per observation in the post-stratification analysis. Most studies, and certainly those of cancer, would match for age and sex, since both could confound the effect of most other factors. A large number of studies match on additional variables, often to the point where each case may be associated with a set of controls in an individual stratum. One purpose of this matching, as we have mentioned, is to improve the precision of the estimates of the relevant relative risks obtained from a stratified analysis. Some matching factors, such as place of residence or membership of a sibship, represent a complex of factors. Then, the purpose of the matching is to eliminate the confounding effect of a range of only vaguely specified variables, since the matching provides a stratification by these variables which would otherwise be difficult to perform because of their indeterminate nature. In these circumstances, matching can be an important way of eliminating bias in the risk estimate. The result given for a dichotomous matching variable can be extended without difficulty to any complexity of matching variables. The expected odds ratio resulting from an analysis incorporating the matching is always more extreme than the expected odds ratio obtained ignoring the matching (Seigel & Greenhouse, 1973).

Now, the purpose of matching implies that the matching factors must *a priori* be considered as ones for which stratification would be necessary, that is, as confounding variables. It would follow *that variables which have been used for matching in the design should be incorporated in the analysis as confounding variables*. Until recently,

there were limitations on the type of analysis that could be done which fully incorporated the matching. However, the analytical methods now available do not suffer from these limitations.

The extent to which the analysis should incorporate the matching variables will depend on how the variables are used for matching. If matching is performed only on age and sex then a stratified analysis rather than one which retains individual matching may be more appropriate. Individual matching in the analysis is only necessary if matching in the design was genuinely at the individual level. However, preservation of individual matching, even if artificial, can sometimes have computational advantages and often means little loss of information (see § 7.6).

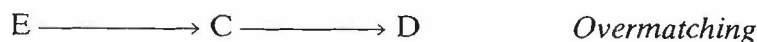
Overmatching

It might be inferred from our discussion that the post-stratification analysis is always the one of interest, that if we can find a variable which appears to alter the association between disease and the exposure then we should treat that variable as a confounder, but this approach ignores the biological meaning of the variables in question and their position in the sequence of events which leads to disease.

A diagrammatic representation of (positive) confounding might be as follows:



In many situations, however, such a figure does not correspond to the true state of affairs. Two such situations merit particular attention. The first is when an apparent confounding variable in fact results from the exposure it appears to confound. We could represent this occurrence diagrammatically as



where C is part of the overall pathway



Chronic cough, smoking, and lung cancer can be cited as an example. One would expect the pattern of cigarette smoking among those with chronic cough to be closer to the smoking pattern of lung cancer cases than to that of the general population. The result of stratifying by the presence of chronic cough before diagnosis of lung cancer might almost eliminate the lung cancer-smoking association. The real association between smoking and lung cancer is obscured by the intervention of an intermediate stage in the disease process. A similar example is given by cancer of the endometrium, use of oestrogens and uterine bleeding. If use of oestrogens by postmenopausal women induces uterine bleeding, itself associated with endometrial cancer, then one might expect stratification by a previous history of uterine bleeding to reduce the association between endometrial cancer and oestrogens. If history of chronic cough or a history of uterine bleeding were used as stratifying factors in the respective analyses, or as a matching factor in the design, then one would call the resulting reductions in the

strength of the disease/exposure association examples of overmatching. In both these examples, the overmatching consists of using as a confounding factor a variable whose presence is caused by the exposure.

A second way in which overmatching may occur is when both the exposure and the confounder represent the same underlying cause of the disease. We might represent such a situation as:



C and E now represent different aspects of the same composite factor causally related to disease. For example, C and E might both be aspects of dietary fibre, or alternative measures of socioeconomic status. From the diagram it is clear that both should have equal status as associates of disease. One might, somewhat arbitrarily, decide to take one of the two, or even attempt to form a composite variable using regression methods. It would clearly be inappropriate to consider one as confounding the effect of the other, or to consider the association of one with disease after stratification by the other.

In both the above situations, overmatching will lead to biased estimates of the relative risk of interest.

A third way in which overmatching may occur is through excessive stratification. The standard errors of post-stratification estimates of relative risk tend to be larger than the standard errors of pre-stratification estimates (see § 7.0). Stratification by factors which are not genuine confounding variables will therefore increase the variability of the estimates without eliminating any bias, and can be regarded as a type of overmatching. It is commonly seen when data are stratified by a variable known to be associated with exposure but not in itself independently related to disease. It does not give rise to bias. If one recalls the section of Chapter 1 relating to overmatching in the design of a study, one can see close parallels between the different manifestations of overmatching in the design and the analysis of a study.

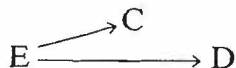
We may represent diagrammatically the situation in which a variable is related to exposure but not to disease as:



C is not a genuine confounding variable. Simply by chance, however, substantial confounding may appear to occur, as the result of random sampling. This eventuality will arise more often the more strongly E and C are associated. It may be difficult to decide whether such an event has occurred, and this will normally require consideration of how C and disease could, logically or biologically, be related. If several studies have been performed, such confounding may appear as an inconsistency in the results, with different factors appearing to have confounding effects in different studies. Good evidence may be available from previous studies that C is not causally related to disease, in which case it should not be incorporated as a confounder. If, nevertheless, it appears

to have a strong confounding effect, the design of the study should be carefully examined to ensure that it is not acting as a surrogate for some other potential confounder, and in particular that it is not acting through selection bias (see § 3.7).

Note that the situation



may lead to genuine confounding when the variables are measured with error. Consider as an example:

	Factor C+ Exposure E		Factor C- Exposure E		Pooled levels of C Exposure E	
	+	-	+	-	+	-
Case	90	1	10	9	100	10
Control	9	10	1	90	10	100
Odds ratio	100		100		100	

Suppose now that E is misclassified 10% of the time, yielding a variable which we shall denote by E*. The tables become, approximately:

	Factor C+ Exposure E*		Factor C- Exposure E*		Pooled levels of C Exposure E*	
	+	-	+	-	+	-
Case	81	10	10	9	91	19
Control	9	10	10	81	19	91
Odds ratio	9		9		22.9	

The post-stratification odds ratio relating E* to disease is much less than that relating E to disease but in addition a confounding effect has arisen, with a confounding risk ratio of $22.9/9 = 2.54$. The odds ratio relating C to disease, after stratification by E*, is 9 rather than 1. The reason is clear: both E* and C are correlates of E, and both are related to disease only through E. Only if E is exactly known does knowledge of C contribute nothing extra to assessment of disease risk.

It is clear from our discussion of confounding that it is not an issue which can be settled on statistical grounds. One has to consider the nature of the variables concerned, and of their relationships with each other and with disease.

Variables to be included as confounding variables

We have considered the conditions under which an observed association may be the result of a confounding effect, and when overmatching might occur, and have discussed the criteria for deciding which factors to incorporate in the analysis as confounding

factors when confronting the data from a particular study. Normally there will be two basic aims: first, to remove from the disease/exposure associations of interest all the confounding effects present in the study data set, whether positive or negative; second, to ensure that genuine associations are not reduced by overmatching.

To satisfy the first aim, questions of statistical significance are irrelevant. Given that a confounding factor has to be associated both with disease and with exposure, one might contemplate testing whether both associations are significant in the available data. If their association were not significant, then one might discard the factor as a potential confounder. *This approach is incorrect* (Dales & Ury, 1978), and it can lead to substantial confounding effects remaining in the association, as the following example shows:

Stratification by potential confounder						
	Factor C+		Factor C-		Pooled levels of C	
	Exposure E		Exposure E		Exposure E	
	+	-	+	-	+	-
Case	80	40	5	5	85	45
Control	8	4	40	40	48	44
Odds ratio	1		1		1.73 ($\chi^2 = 3.91$)	

The association between E and C, after cross-classification by disease status, does not achieve significance at the 5% level: $\chi^2 = 1.63$ using the Mantel-Haenszel χ^2 given by equation (4.23). Thus, in the data C and E are not significantly associated, but an appreciable and statistically significant (at least in the formal sense) association between E and disease exists before stratification by C, which vanishes upon stratification.

With this example in mind, and recalling the initial discussion of overmatching, we can propose three criteria for treating a variable as a confounding variable in the analysis.

1. If a variable C is known from other studies to be related to disease, and if this association is not subsidiary to a possible exposure/disease association, then C should be treated as a confounding variable. The significance of the association between C and disease in the data at hand is of no relevance. Irrespective of the association between E and C in the general population, if there is an association between E and C in the study sample then part of the association between E and disease in the study sample will be a reflection of the causal association between C and disease. The contribution of C to the E-disease association must be eliminated before proceeding to further considerations of a possible causal role for E in disease development. Age and sex will almost always be confounding variables, and should be treated as such.
2. If a variable C is related to disease, but this association is subsidiary to the association between E and disease, by which we mean that either C is caused by E or forms a part of the chain of events by which disease develops from E, then C should not be considered as a confounder of the disease/exposure association.

3. If a factor is thought important enough to be incorporated in the design of the study as a matching or balancing factor, then it should be treated as a confounding variable in the analysis.

In the situation when E and C are known to be related, and if in the data C is also related to disease, then there will be an apparent confounding effect. In this situation, unlike the previous one, it is less clear what the interpretation should be in terms of causality. Incorporating C as a confounding variable implies that one is giving the C-disease association precedence over the E-disease association, which one would not always want to do, as for example when C and E are different measures of the same composite factor. The possibility must be considered that selection bias has operated with respect to C in the choice of either cases or controls or in the manner of acquiring information. Control of this bias may be possible by treating it as if it were a confounding effect. This is discussed in § 3.7.

3.5 Interaction and effect modification

In our discussion of the joint effect of different factors, and specifically in the context of confounding, we have assumed that the odds ratio associating one factor to disease is unaltered by variation in the value of other factors. This simple assumption can only be an approximation, although as we saw in Chapter 2, on many occasions the approximation is fairly close. On other occasions, appreciable variations in the odds ratio were noted, and these variations themselves were of biological importance.

If the odds ratios associating factor A and disease vary with the level of a second factor B, then it is common epidemiological parlance to describe B as an effect modifier. The term is not a particularly happy one, however. A departure from a multiplicative model might arise, for example, if two factors operated in the same way at the cellular level and their joint effect were additive, which would make little sense biologically to describe as effect modification. We prefer to use the term 'interaction', in keeping with usual statistical terminology.

The main reasons for studying interactions are first because they may modify the definition of high risk groups, and second because they may provide insight into disease mechanisms. Interaction implies that in certain subgroups the relative risk associated with exposure is higher than in the rest of the population. Both the specificity of risk for these subgroups, and the fact that the level of exposure-associated risk will be higher than the general risk in the population would tend to increase one's belief in the causal nature of the association, as was discussed earlier in the chapter. The aim should not be to eliminate interactions by suitable transformations, but rather to understand their nature; this point is well made by Rothman (1974).

One should note that using a variable as a matching factor in the design, so that its individual effect on risk cannot be studied, does not alter the interactive effects that the factor may have with other exposures. A simple example will illustrate the point.